A Tutorial for IMDB-Database-Explorer

Introduction

IMDB-Database-Explorer is a browser-based R Shiny app for exploring the datasets in IMDb website and finding out the relationship between different variables in the datasets. In this app, two datasets from IMDb website have been used, one is title basics and the other one is title ratings. Variables used from these two datasets in the app are titleType, isAdult, startYear, runtimeMinutes, genre from titles basic dataset, and averageRating and numVotes from title ratings dataset. After downloading data from the IMDb webpage site (https://datasets.imdbws.com/), data processing (to add two datasets in one single file, remove NA) has been conducted to have the final dataset for the app. The data is available at github page: https://github.com/tasnimmajumder/IMDB-Database-Exploration/tree/main/data. The IMDB-Database-Explorer is publicly available at https://github.com/tasnimmajumder/IMDB-Database-Exploration. This tutorial can also be downloaded from the github page: https://github.com/tasnimmajumder/IMDB-Database-Exploration/blob/main/IMDB-Database-Explorer%20Tutorial.pdf. Specification of each tab of IMDB-Database-Explorer app will be described in the following sections.

How to Start

This is an instruction of how to install and run IMDB-Database-Explorer shiny software locally (https://github.com/tasnimmajumder/IMDB-Database-Exploration).

Requirement:

• R (>= 4.0.2)

• Shiny (>=1.2.0)

How to install shiny package:

1. Open R.

2. User can install the shiny package by the following command in R:

install.packages("shiny ")

install.packages("shinythemes ")

How to install and run IMDB-Database Explorer locally

1. Open R.

2. Run IMDB-Database Explorer by the following commands in R:

library(shiny)

library(shinythemes)

shiny::runGitHub("IMDB-Database-Exploration", "tasnimmajumder", ref="main")

(The first tab (Descriptive Statistics) of IMDB-Database-Explorer app will pop-up, **Figure 1**)

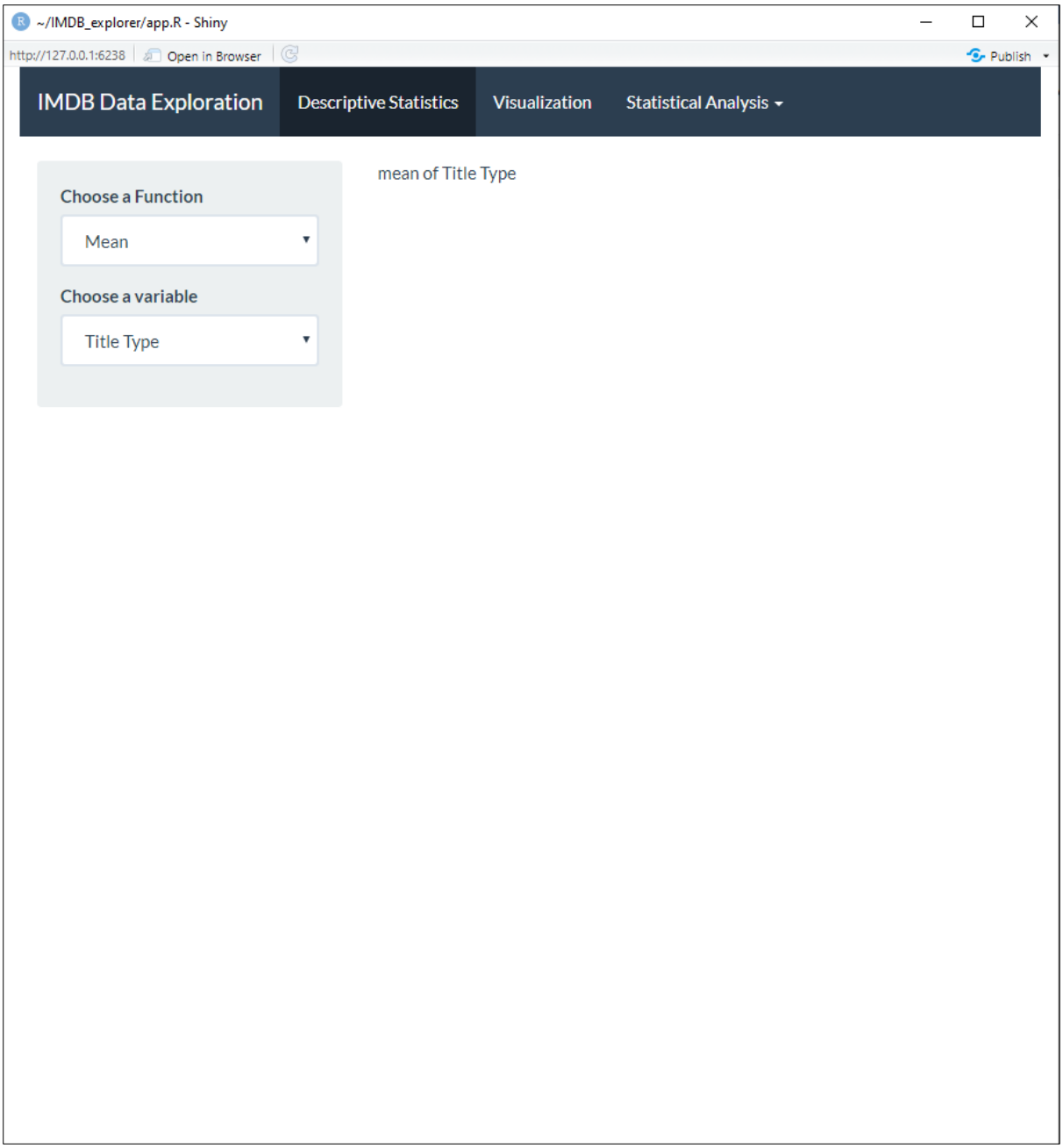

Figure 1: IMDB-Database-Explorer App initially

Descriptive Statistics (First Tab)

In this tab descriptive statistics for the variables in final dataset can be done. For Continuous variables, Start Year (startYear), Average Rating (averageRating), Runtime in Minutes (runtimeMinutes) and Number of Votes (numVotes), following functions are provided: mean, median, standard deviation, maximum and minimum. For categorical variables, Title Type (titleType), Genre (genre) and Is Adult (isAdult), summary table function has been provided.

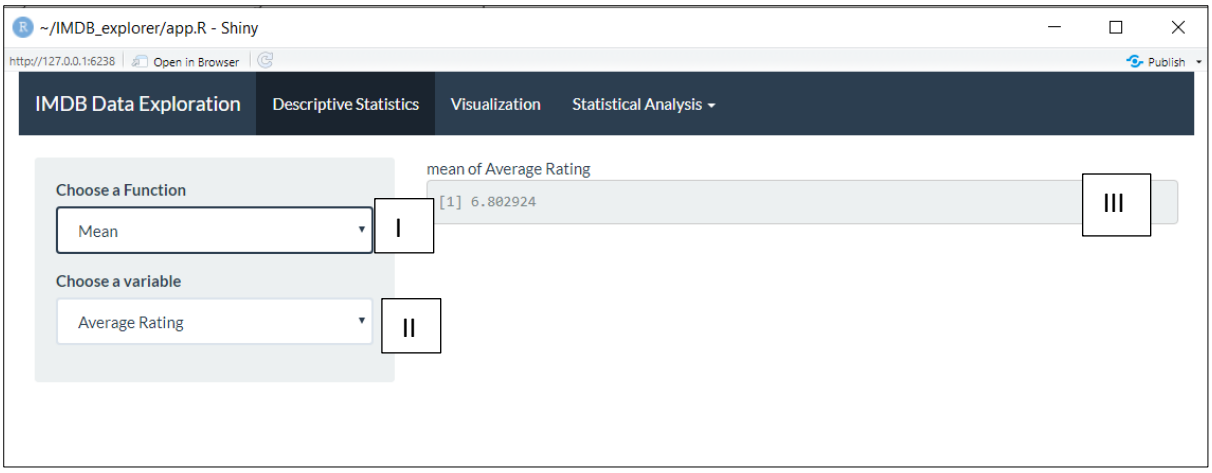

Figure 2: Descriptive statistics of continuous variable

In descriptive statistics tab, user can select function (Figure2-I) and choose variable (Figure 2-II). Outcome of the selected for choosing variable will be shown in the main panel of the tab (Figure 2-III). For continuous variables, Mean, Median, standard Deviation, Maximum and Minimum function will work and using these functions will get the mean value, median value, standard deviation of the variable, maximum value and minimum value of the selected continuous variable. In figure 2, it is depicted that selected function is Mean and selected variable is Average Rating and the mean value of average rating is 6.802924.

For categorical variable, summary table function will provide descriptive statistics. It will provide the number of observations for each category within the variables. In figure 3, we can see that selected function is Summary Table and selected variable is Genre, and in the main panel number of observations in the dataset for each 'Genre' category is shown as a table.

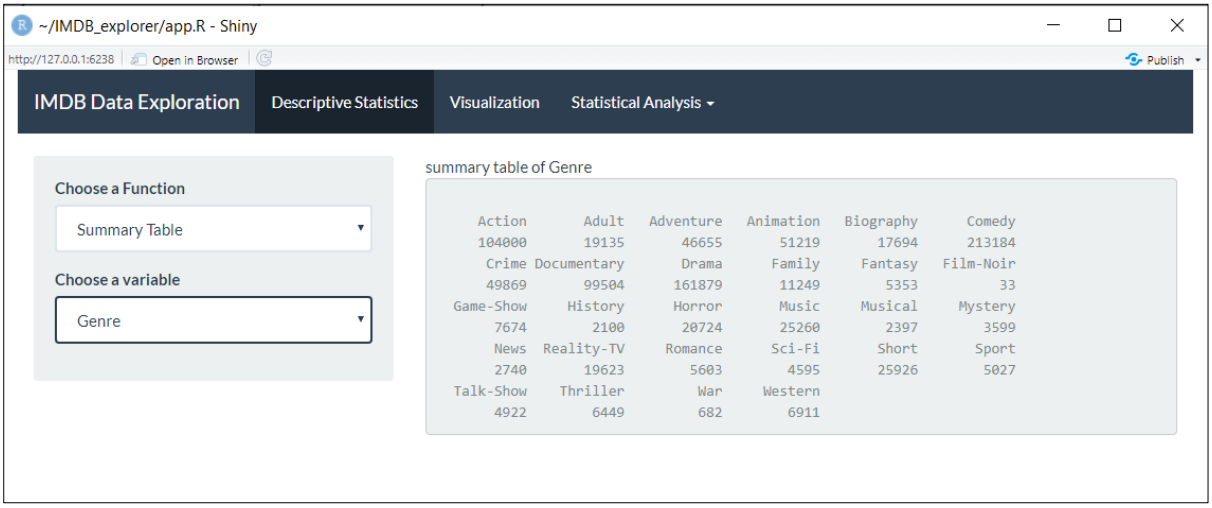

Figure 3: Descriptive statistics for categorical variables

Visualization (Second Tab)

In the visualization tab, user can investigate the relationship between variables in IMDb dataset. For showing the relationship among two continuous variables and one categorical variable geom_smooth and geom_point have been used. On the other hand, for depicting the relationship among two categorical variables and one continuous variable geom_boxplot has been used. To show the relationship among the variables, final dataset has been filtered for running time less than 300 minutes for removing the outliers. In most cases, the duration of program for any Title Type or Genre can not be more then 300 minutes (5 hour).

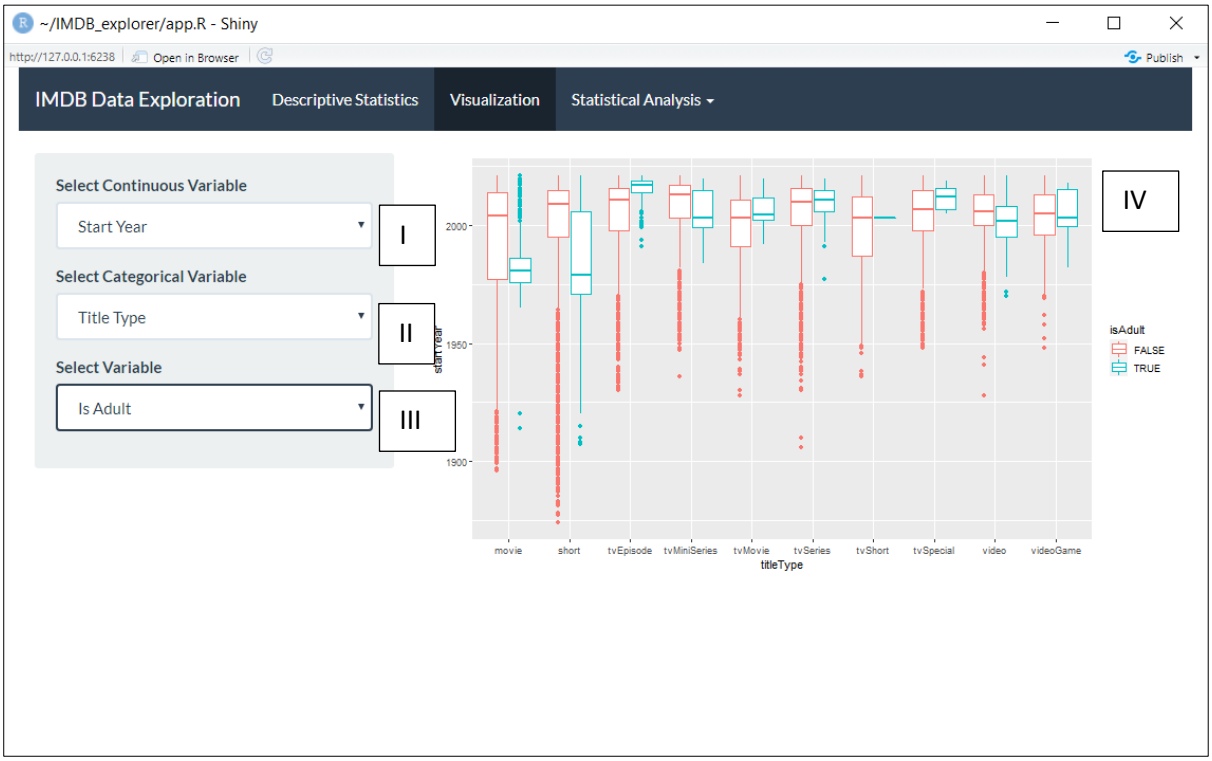

Figure 4: Visualization tab for showing relationship among the variables

For showing the relationship among there variables, user can select one continuous variable (Figure 4- I), one categorical variable (Figure 4- II) and any other continuous or categorical variable (Figure 4- III). For selecting two categorical variables and one continuous variable a boxplot will appear in the main panel of visualization tab (Figure 4- IV). The boxplot depicts that for each Title Type, Start Year distribution for adult (TRUE) and non-adult (FALSE) title.

For two continuous variables and one categorical variable, the graph appears like the following figure (Figure 5).

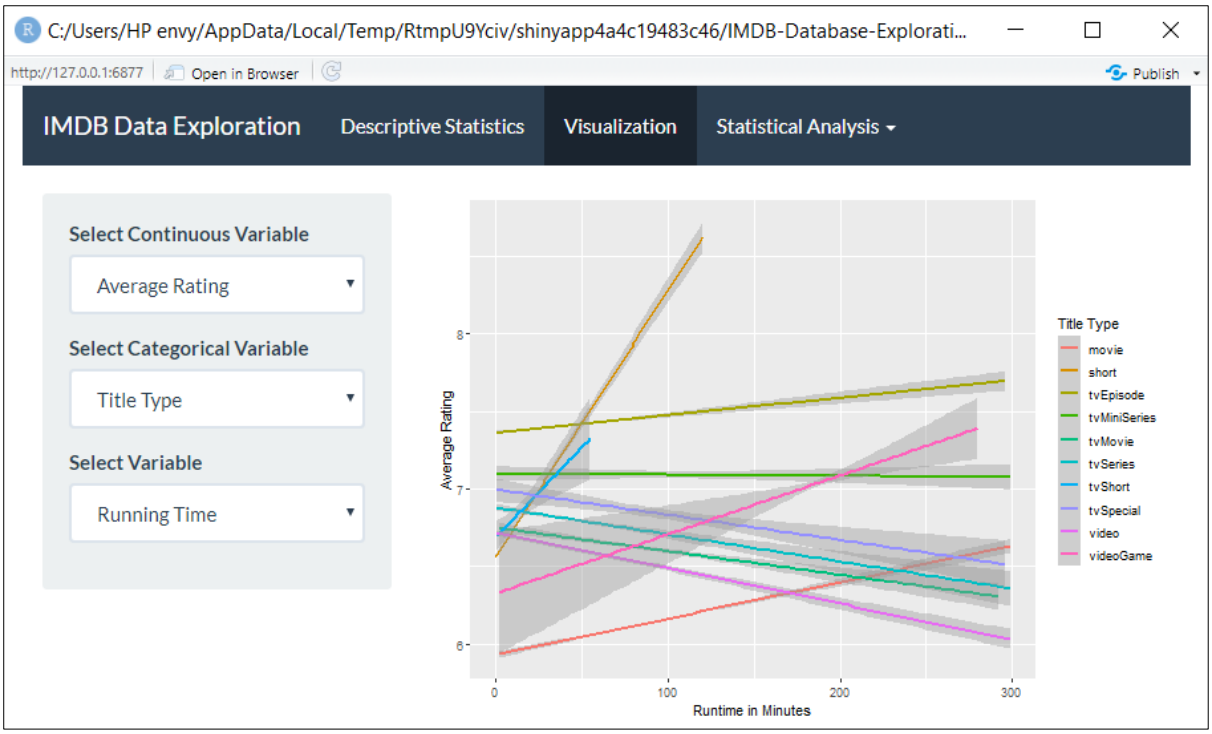

Figure 5: Relationship among average rating and running time for each title type

From the graph we can see that how average rating differs with increasing running time for each title type. For example, in case of ‘movie’ Title Type as running time increases average rating will be increased from 6 to 6.5.

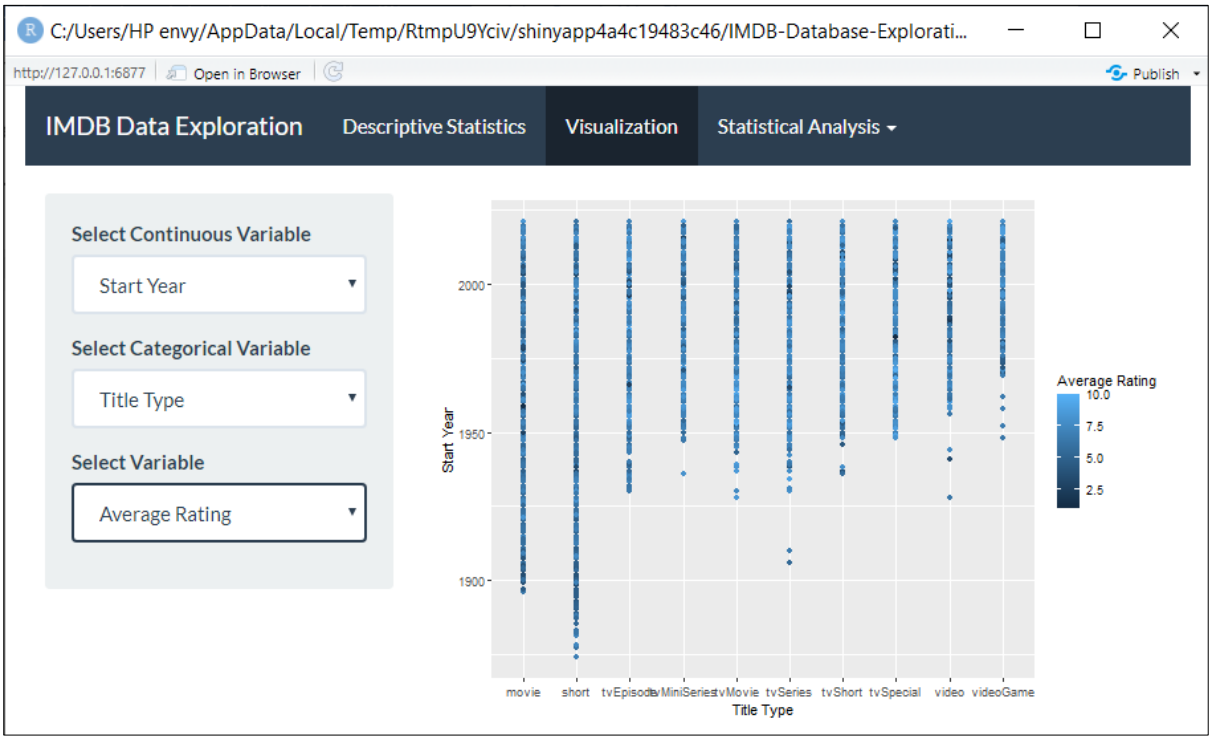

Figure 6: Relationship among start year, title type and average rating

In Figure 6, it shows relationship among start year, title type and average rating variables. It depicts for each Title Type how average rating has changed by each start year.

Statistical Analysis (Tab 3)

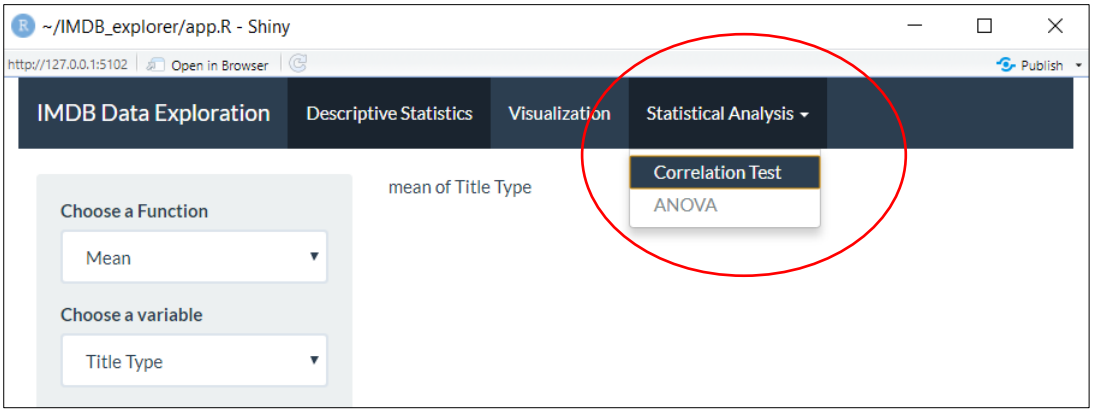

Figure 7: Statistical analysis tab

In the third tab, Statistical Analysis, two statistical tests can be performed. One is correlation test for hypothesis testing between two continuous variables. Other one is ANOVA test for hypothesis testing between one continuous and one categorical variable. For performing the statistical analysis 'Strat Year' variable is not considered because considering it has no statistical influence on other variables of the dataset and final dataset has been filtered for running time less than 300 minutes for removing the outliers.

For 'Correlation Test' user have to select two continuous variables (Figure 8- I and II) and the output of the test will be shown in the main panel of the tab (Figure 8- III). Correlation value between two variables will be found in Figure 8- V and the p-value for checking whether the correlation value is statistically significant will be found in Figure 8- IV.

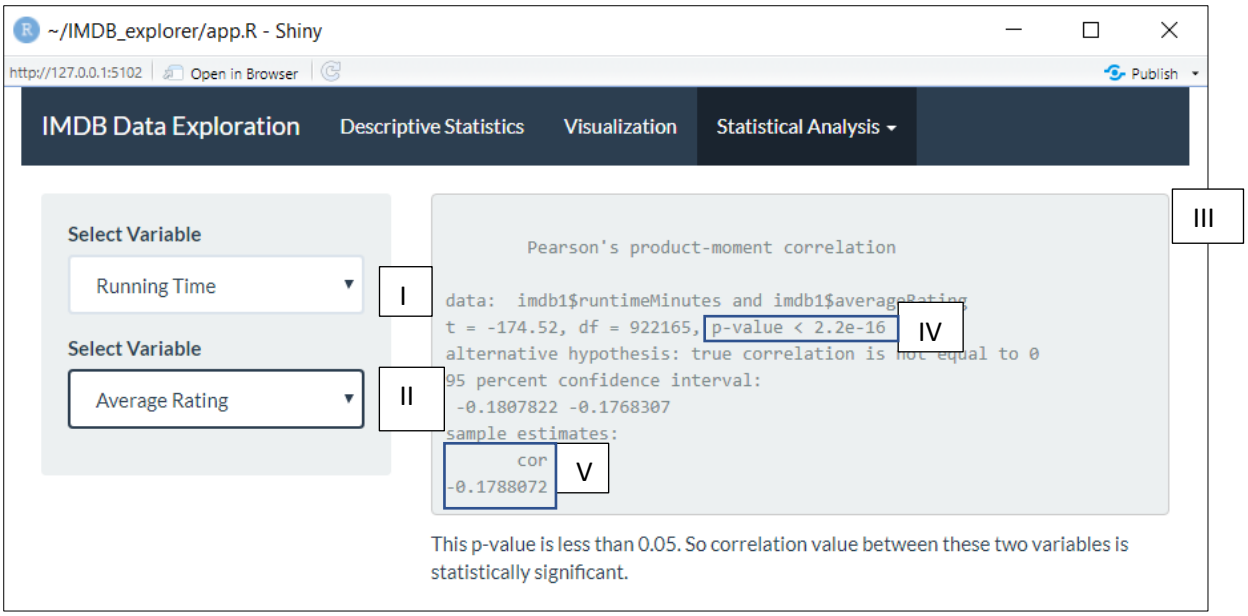

Figure 8: Correlation test between two continuous variables

In the above figure, we can see the output of correlation test between Running Time and Average Rating variable. Correlation value is -0.1788072 which means that these two variables are negatively co-related. The p-value for correlation test is less than 0.05 which means that correlation value of these two variables is statistically significant.

For 'ANOVA Test' user have to select one continuous variable (Figure 9- I) and one categorical variable (Figure 9- II) and the output of the test will be shown in the main panel of the tab (Figure 9- III). F-value of the ANOVA test will be found in Figure 9- IV and the p-value for checking whether there is any difference in average of continuous variable among the categories of categorical variable correlation will be found in Figure 9- V.

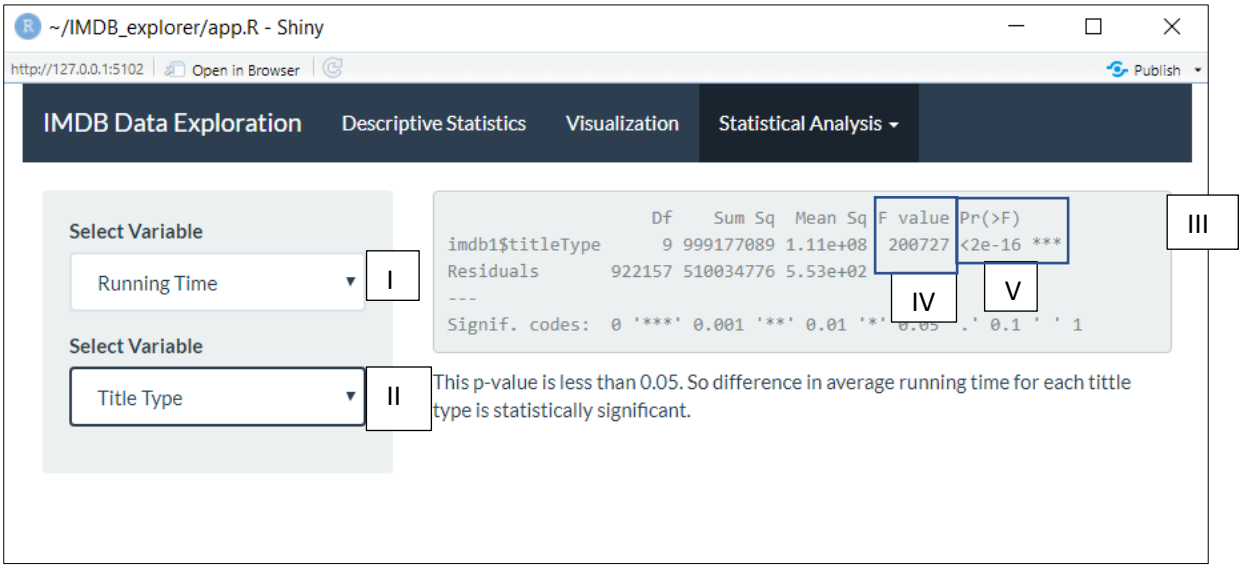

Figure 9: ANOVA test for one continuous variable and one categorical variable

In the above figure, we can see the output of ANOVA test between Running Time and Title Type variable. F-value of the test is very large which means that there is huge variation in running time among the categories of Title Type. The p-value of ANOVA test is less than 0.05 which means that difference in average running time for each Title Type is statistically significant.